An Adaptive Directional Metropolis-within-Gibbs algorithm

Yan Bai*

April 6, 2009

Abstract

In this paper we propose a simple adaptive Metropolis-within-Gibbs algorithm attempting to study directions on which the Metropolis algorithm can be ran flexibly. The algorithm avoids the wasting moves in wrong directions by proposals from the full dimensional adaptive Metropolis algorithm. We also prove its ergodicity, and test it on a Gaussian Needle example and a real-life Case-Cohort study with competing risks. For the Cohort study, we describe an extensive version of Competing Risks Regression model, define censor variables for competing risks, and then apply the algorithm to estimate coefficients based on the posterior distribution.

1 Introduction

Markov Chain Monte Carlo methods (MCMC) are used to do simulations based on constructing Markov Chain, and widely applied for physics, statistics, biology, genetics, cryptography and others. One inspiring algorithm was proposed by Metropolis et al. (1953). Through a symmetric proposal transition, the Metropolis algorithm prescribes a transition rule for Markov Chain. Hastings (1970) later generalized it to the case that the proposal distribution is not necessarily symmetric.

Adaptive algorithms are generally used to learn parameter information of target distribution from historical sample data. Some adaptive MCMC methods using regeneration times and other complicated constructions had been proposed by Gilks et al. (1998); Brockwell

*Department of Statistics, University of Toronto, Toronto, ON M5S 3G3, CA. yanbai@utstat.toronto.edu

and Kadane (2005), and elsewhere. Haario et al. (2001) proposed an adaptive Metropolis algorithm attempting to improve the convergence, and proved its ergodicity. Following that, many general results were developed, see Andrieu and Robert (2002); Atchadé and Rosenthal (2005); Andrieu and Moulines (2006); Roberts and Rosenthal (2007); Yang (2008); Saksman and Vihola (2008); Bai et al. (2008); Atchadé and Fort (2008); Craiu et al. (2008); Bai (2009).

In Section 2 we review Metropolis-Hastings algorithm, Metropolis-within-Gibbs sampler, and certain adaptive Metropolis algorithm. Their common ground is based on constructing reversible Markov Chain in each step. Although these algorithms are very successful for many target distributions, they can not work efficiently for some cases that either many wasting jumps are generated, or many moves with small sizes are generated by reason of the limitation from the jumping directions. The phenomenon is very explicit for high dimensional cases or some low dimensional extreme cases. A toy example will be presented in Section 3 for explanations.

In Section 4 we propose the adaptive algorithm: *adaptive Directional Metropolis-within-Gibbs* (ADMG) algorithm which can avoid the problem. The idea is similar as that of the Hit-and-Run algorithm. The framework of Hit-and-Run is to uniformly draw a random direction in the unit hypersphere, and then sample a scalar from some proposal distribution on the chosen direction, see literatures Bélisle et al. (1993); Chen and Schmeiser (1993); Gilks et al. (1994); Roberts and Gilks (1994); Chen and Schmeiser (1996); Kaufman and Smith (1998); Lovász (1999); Lovász and Vempala (2003, 2006); Bèdard and Fraser (2008). Metropolis with single particle moves, Gibbs sampler, Swendsen-Wang, data augmentation, and slice sampling have the same basic structure, see Andersen and Diaconis (2007). The ADMG algorithm tries to find directions and corresponding jumping scalars through studying certain estimate of empirical covariance matrix of the sample chain. Then Metropolis-within-Gibbs sampler is ran on the seized directions with the jumping scalars as variances. The method can suppress the proportion of wasting moves by proposals from full dimensional Metropolis algorithm. We also compare it with Metropolis-within-Gibbs sampler and adaptive Metropolis algorithm through analysing the toy example on 10-dimensional Euclidean space. Then we show its ergodicity.

In Section 5 we discuss a real-life Case-Cohort study for the application, where the dataset was from the Princess Margaret Hospital, a leading cancer centre in North America. Cohort study is commonly based on the survival model. In practice, the likelihood function turns to be more and more complicated as the number of observations increases. The trade-off alternative, partial likelihood function is more interesting. Given a prior distribution, we consider the posterior distribution, and implement our algorithm to find the estimate of the coefficients of the interest covariates in the study.

2 Background

Let the state space $\mathcal{X}$ be an open set in $\mathbb{R}^d$ with Borel σ-field $\mathcal{F}$ and target density $t : \mathcal{X} \longrightarrow (0, \infty)$ with $\int_{\mathcal{X}} t(x)\mu(dx) < \infty$ where μ is d-dimensional Lebesgue measure. Given the X_n, the *Metropolis-Hastings* algorithm generates the proposal $Y_{n+1} \sim Q(X_n, \cdot)$ with the measurable density function $q : \mathcal{X} \times \mathcal{X} \longrightarrow [0, \infty)$. Let

$$\alpha(x, y) := 1 \wedge \frac{t(y)q(y, x)}{t(x)q(x, y)}. \tag{2.1}$$

Then X_{n+1} is assigned Y_{n+1} with the probability $\alpha(X_n, Y_{n+1})$, and is assigned X_n with the probability $1 - \alpha(X_n, Y_{n+1})$. If $q(x, y) = q(y, x)$, call it *Metropolis algorithm*.

Roberts and Rosenthal (2006b) studied the conditions under which the *Metropolis-within-Gibbs* (MG) algorithm is Harris recurrent or not. Fort et al. (2003) presented some conditions under which the random-walk-based Metropolis-within-algorithm is geometrically ergodic. Roberts and Rosenthal (2006a) studied certain adaptive Metropolis-within-Gibbs algorithm for the hierarchical model.

For $1 \leq i \leq d$, let $q_i : \mathcal{X} \times \mathbb{R} \longrightarrow [0, \infty)$ be jointly measurable with $\int_{-\infty}^{\infty} q_i(x, z)dz = 1$ for all $x \in \mathcal{X}$ where dz is one dimensional Lebesgue measure. Let $Q_i(x, \cdot)$ be the Markov kernel on $\mathbb{R}^d$ which replaces the ith coordinate by a draw from the density $q_i(x, \cdot)$, but leaves the other coordinates unchanged. That is

$$Q_i(x, \mathcal{S}_{i,a,b}) = \int_a^b q_i(x, z)dz, \tag{2.2}$$

where

$$\mathcal{S}_{i,a,b} := \{y \in \mathcal{X} : y_j = x_j \text{ for } j \neq i \text{ and } y_i \in [a, b]\}.$$

Say $Q_i(x, \cdot)$ is symmetric if

$$q_i((x_1, \cdots, x_d), z) = q_i((x_1, \cdots, x_{i-1}, z, x_{i+1}, \cdots, x_d), x_i).$$

For $x, y \in \mathbb{R}^d$ and $1 \leq i \leq d$, let

$$\alpha_i(x, y) := \mathbf{1}\left(t(x)q_i(x, y_i) \neq 0\right) \min\left[1, \frac{t(y)q_i(y, x_i)}{t(x)q_i(x, y_i)}\right] + \mathbf{1}\left(t(x)q_i(x, y_i) = 0\right). \tag{2.3}$$

Let P_i be the kernel which proceeds as follows. Given X_n, it generates the proposal $Y_{n+1} \sim Q_i(X_n, \cdot)$. Then X_{n+1} is assigned Y_{n+1} with the probability $\alpha_i(X_n, Y_{n+1})$, and is assigned X_n with the probability $1 - \alpha_i(X_n, Y_{n+1})$.

Let I_n be a random variable on $\{1, \cdots, d\}$. Two most common schemes are *deterministic-scan Metropolis-within-Gibbs sampler* $P_{DS} = P_{I_n}$ where $I_n = n \mod d$, and *random-scan*

Metropolis-within-Gibbs sampler $P_{RS} = P_{I_n}$ where I_n is uniform on $\{1, \cdots, d\}$. Then for $n = 0, 1, 2, \cdots$ given X_n and I_n the chain $X_{n+1} \sim P_{I_n}(X_n, \cdot)$. It is straightforward to verify that the chain has stationary distribution $\pi(\cdot)$ given by

$$\pi(A) := \frac{\int_A t(x)\mu(dx)}{\int t(x)\mu(dx)}, \qquad A \in \mathcal{F}.$$

For the above two algorithms, at each step, the proposal distribution is fixed (independent of the time n and historical information) so the sampled chain $\{X_n\}$ is a time-homogeneous Markov Chain.

Now we describe the *adaptive MCMC* algorithm. Let $\{P_\gamma : \gamma \in \mathcal{Y}\}$ be a collection of Markov Chain kernels on the state space $\mathcal{X}$. $\mathcal{X}_n \in \mathcal{X}$ and $\Gamma_n \in \mathcal{Y}$ are respectively the sample point and the random kernel index at the time n. Γ_n is chosen according to adaptation scheme, and actually it is a function of X_n and $\{(X_i, \Gamma_i) : i = 0, 1, \cdots, n-1\}$. Given X_n and Γ_n, X_{n+1} is generated from the random kernel $P_{\Gamma_n}(X_n, \cdot)$. Say the adaptive MCMC algorithm (X_n, Γ_n) is *ergodic* if for any initial point $(x_0, \gamma_0) \in \mathcal{X} \times \mathcal{Y}$, the distance between $P(X_n \in \cdot \mid X_0 = x_0, \Gamma_0 = \gamma_0)$ and the target distribution $\pi(\cdot)$ converges to zero under the total variation norm.

Roberts and Rosenthal (2006a) introduced an *adaptive Metropolis* (AM) algorithm which is a slight variant of the algorithm of Haario et al. (2001). At the n^{th} iteration, the proposal distribution $Q_n(x, \cdot) = N(x, 0.1^2\mathbf{1}_d/d)$ for $n \le 2d$; for $n > 2d$,

$$Q_n(x, \cdot) = \begin{cases} (1-\theta)N(x, (2.38)^2\Sigma_n/d) + \theta N(x, (0.1)^2 I_d/d), & \Sigma_n \text{ is positive definite,} \\ N(x, (0.1)^2 I_d/d), & \Sigma_n \text{ is not positive definite,} \end{cases} \tag{2.4}$$

for some fixed $\theta \in (0, 1)$, and the empirical covariance matrix

$$\Sigma_n = \frac{1}{n}\left(\sum_{i=0}^{n} X_i X_i^\top - (n+1)\overline{X}_n \overline{X}_n^\top\right), \tag{2.5}$$

where $\overline{X}_n = \frac{1}{n+1}\sum_{i=0}^{n} X_i$, is the current modified empirical estimate of the covariance structure of the target distribution based on the run so far. Commonly, the iterative form of Equation (2.5) is more useful,

$$\Sigma_n = \frac{n-1}{n}\Sigma_{n-1} + \frac{1}{n+1}\left(X_n - \bar{X}_{n-1}\right)\left(X_n - \bar{X}_{n-1}\right)^\top. \tag{2.6}$$

3 A Toy Example

Let the target density

$$t(x) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-x^\top \left(\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \right)^{-1} x/2 \right), \tag{3.7}$$

where $\theta = 45^o$, $\sigma_1 = \sqrt{20}$ and $\sigma_2 = 0.01$. The target distribution has extremely small variance 0.0001 and large variance 20 respectively on the two directions $(-\sqrt{2}/2, \sqrt{2}/2)$ and $(\sqrt{2}/2, \sqrt{2}/2)$. So, the target is mainly supported on a very narrow region along the 45° degree direction between the x_1-axis and the x_2-axis. The length of the needle region is roughly $2 * 4 * \sqrt{20} = 35.78$ (see the true sample data in Figure 4.2) because $P(|Z| < 4) \approx 1$ where Z is standard normal. We run MG sampler and AM to generate target sample data with the same initial point $X_0 \sim N(\overrightarrow{0}, diag(1,1))$. However, the results are not satisfying.

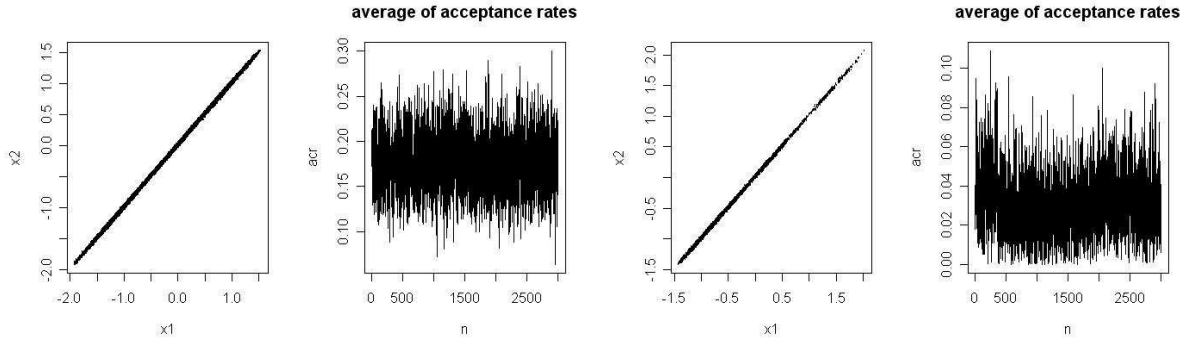

Figure 3.1: The first plot is the sample plot by running random-scan MG sampler. The second plot is the 100-step average of acceptance rates by random-scan MG sampler. The third plot is the sample plot by running AM. The last plot is the 100-step average of acceptance rates by AM algorithm.

Given the sampled data $\{X_0, X_1, \cdots\}$ and the proposal values $\{Y_1, Y_2, \cdots\}$, the *k-step average of acceptance rates* is defined as

$$\alpha_i^{(k)} := \frac{1}{k} \sum_{t=ki}^{k(i+1)-1} \alpha(X_t, Y_{t+1}), \tag{3.8}$$

where $i = 0, 1, \cdots$.

We perform the random-scan MG sampler by $300,000$ iterations using the normal distribution with variance 0.1 as the proposal distribution, see the left two plots of Figure 3.1. From

the sample plot, the sample data has the needle shape with the length around $4.95 \ll 35.78$ roughly between the two points, $(-2.0, -2.0)$ and $(1.5, 1.5)$. The 100-step average $\{\alpha_n^{(100)}\}$ of acceptance rates is roughly between 0.10 to 0.3. We also tried normal proposals with different variances 0.0001 (same as the target's) which also gives worse results. For random-scan MG sampler, at each step, the jumping direction of the sample chain can be just either on the axis x_1 or the axis x_2 so the jumping scale is strongly limited. Moreover, the 100-step average of acceptance rates is very sensitive to the proposal variance. When the proposal variance is large, the proposal values are easily rejected. When the proposal variance is small, the proposal values are easily accepted but the chain is easily stuck.

We also perform by $300,000$ iterations AM stated in Section 2. The algorithm attempts to find a better transition kernel by learning the empirical covariance matrix Σ of the sample chain. The sample points also span roughly the narrow stripe with the length around $4.95 \ll 35.78$ between the two points, $(-1.5, -1.5)$ and $(2, 2)$, see the third plot in Figure 3.1. At the same time, the 100-step of acceptance rates is quite small, see the last plot in Figure 3.1. So the sampling method for this example also does not work well.

To find the reason of the inefficiency of AM, let us observe the estimate of empirical covariance matrix Σ_n for $n = 300,000$,

$$\Sigma_n = \begin{bmatrix} 1.449585 & 1.448932 \\ 1.448932 & 1.449020 \end{bmatrix}.$$

By singular value decomposition, we have $\Sigma_n = UDV$ where

$$\begin{aligned} U &= \begin{bmatrix} -0.7071758 & -0.7070378 \\ -0.7070378 & 0.7071758 \end{bmatrix}, \\ D &= \begin{bmatrix} 2.8982345593 & 0 \\ 0 & 0.0003700081 \end{bmatrix}, \\ V &= U^\top. \end{aligned} \tag{3.9}$$

It is not difficult to find that the matrix U is approximately equal to the U matrix by singular value decomposition on the true covariance matrix of the Gaussian density $t(\cdot)$. The first diagonal element d_1 of D underestimates the variance 20 on the direction $(\sqrt{2}/2, \sqrt{2}/2)$, and the second element d_2 overestimates the variance 0.0001 on the direction $(-\sqrt{2}/2, \sqrt{2}/2)$, see Equation (3.9).

The above fact discloses that AM hardly touches the pinpoint of the needle, actually taking too much time to wander around the middle region of the needle. Only if the jumping direction by AM's proposal is little off the 45° direction between x_1 and x_2, the proposal with a little big jumping scale will be easily rejected. From the last plot in Figure 3.1, the point can be also observed. The 100-step average $\{\alpha_n^{(100)}\}$ of acceptance rates is very low,

approximately below 0.10 that the adaptation finds too much wasting loads of proposal in wrong directions. Hence the inefficiency of AM is mainly due to the jumping directions.

4 The Algorithm and Ergodicity

Drawing a proposal value in the high dimension space involves the direction choice and the jump scale on the direction. The direction choice can be viewed as taking an unit vector on the unit sphere. The jump scale can be viewed as the variance of the proposal marginal distribution on the chosen direction. The aim in ADMG is to find the random directions in which the efficient movement can be ensured. As illustrated in Section 3, the random direction can be withdrawn from the estimate of empirical covariance matrix. After singular value decomposition, the orthogonal transformation can be obtained. Moreover, the diagonal matrix also approximately estimates the target extents in those directions after the rotation. Based on the orthogonal transformation and the extents on the new coordinates, the Metropolis-within-Gibbs sampler can be ran flexibly.

4.1 ADMG

Now we give the notations for Metropolis algorithm on any direction $e \in S^{d-1}$ where S^{d-1} be the unit hypersphere in $\mathbb{R}^d$. For the vector e, $q_e : \mathcal{X} \times \mathcal{X} \longrightarrow [0, \infty)$ be jointly measurable with $\int_{-\infty}^{\infty} q_e(x, ze)dz = 1$ where the integral is along the direction e from $-\infty$ to ∞ and $\langle \cdot, \cdot \rangle$ be inner product on $\mathbb{R}^d$. Let $Q_e(x, \cdot)$ be the Markov transition kernel on $\mathbb{R}^d$ which update the quantity on the direction e by draw from the density $q_e(x, \cdot e)$, but leaves the quantities on other orthogonal directions unchanged. That is

$$Q_e(x, \mathcal{S}_{e,a,b}) = \int_a^b q_e(x, ze)dz, \tag{4.10}$$

where

$$\mathcal{S}_{e,a,b} := \{y \in \mathcal{X} : \langle y, u \rangle = \langle x, u \rangle \text{ for } u \in S^{d-1},\ \langle e, u \rangle = 0 \text{ and } \langle y, e \rangle \in [a, b]\}.$$

For $x, y \in \mathbb{R}^d$ and $e \in S^{d-1}$, let

$$\alpha_e(x, y) := \mathbf{1}\left(t(x)q_e(x, \langle y, e \rangle e) \neq 0\right) \min\left[1, \frac{t(y)q_e(y, \langle x, e \rangle e)}{t(x)q_e(x, \langle y, e \rangle e)}\right] + \mathbf{1}\left(t(x)q_e(x, \langle y, e \rangle e) = 0\right). \tag{4.11}$$

In the paper, we assume that $Q_e(x, \cdot)$ is symmetric i.e., $q_e(x, ae) = q_e(x, -ae)$ for $a \in \mathbb{R}$.

Let P_e be the transition kernel as follows. Given X_n, it generates a proposal $Y_{n+1} \sim Q_e(X_n, \cdot)$. Then with the probability $\alpha_e(X_n, Y_{n+1})$, it accepts the proposal and set $X_{n+1} = Y_{n+1}$; otherwise with the probability $1 - \alpha_e(X_n, Y_{n+1})$ rejects the proposal and set $X_{n+1} = X_n$.

Step 1. Given $X_0, \cdots, X_n$, we can compute the empirical covariance matrix Σ_n defined in Equation (2.5) and (2.6).

Step 2. If the Σ_n is singular, then perform MG and go to Step 6, otherwise run Step 3;

Step 3. Do singular value decomposition: $\Sigma_n = U^{(n)} D^{(n)} V^{(n)}$ where $D^{(n)} := \text{diag}(d_1^{(n)}, \cdots, d_d^{(n)})$, and $U^{(n)}$ and $V^{(n)} = \left(U^{(n)}\right)^\top$ are orthonormal;

Step 4. Compute the random direction $E_i^{(n)} := U^{(n)} e_i$ where $e_i = \underbrace{(0, \cdots, 0, 1, 0, \cdots, 0)}_{i^{th}}$;

Step 5. Perform MG algorithm on the new coordinates $\left(E_1^{(n)}, \cdots, E_d^{(n)}\right)$, where on each direction the proposal distribution is $Q_{E_i^{(n)}}(X_n, \cdot E_i^{(n)})$ (its variance is equal to 0.01 plus $d_i^{(n)} \theta^{(n)}$ where $\alpha_{[n/k]}^{(k)}$ is the average acceptance rate and $\theta^{(n)} = \exp(2d(\alpha_{[n/k]}^{(k)} - 0.3)))$ where k is the number of steps used to calculate the k-step average of acceptance rates, see Equation (3.8);

Step 6. $n := n + 1$ go to 1.

Remark 1. *In step 3, it may takes much times to do singular value decomposition when the state space is high dimensional. However, it is unnecessary to run the computation for each step. The alternative is to do singular value decomposition each m steps. Another method is to only count the accepted sample point to compute the estimate of empirical covariance matrix.*

Remark 2. *In step 5, MG sampler is performed under the rotated coordinates. Either deterministic-scan or random-scan MG sampler can be implemented here. We call the ADMG algorithm using deterministic-scan MG at each step, adaptive directional system-scan Metropolis-within-Gibbs algorithm (ADSSMG), and call ADMG using random-scan MG at each step, adaptive directional random-scan Metropolis-within-Gibbs algorithm (ADRSMG).*

At the n^{th} iteration, the transition kernels for ADRSMG and ADSSMG are respectively

$$P_{DRS,\Gamma_n}(X_n, \cdot) := \frac{1}{d} \sum_{i=1}^{d} P_{E_i^{(n)}}(X_n, \cdot), \text{ and } P_{DSS,\Gamma_n}(X_n, \cdot) := \prod_{i=1}^{d} P_{E_i^{(n)}}(X_n, \cdot), \tag{4.12}$$

where $P_{E_i^{(n)}}$ is the transition kernel derived from the Metropolis-Hastings proposal $Q_{E_i^{(n)}}$ on the direction $E_i^{(n)}$. The direction $E_i^{(n)}$ is a function of Σ_n depending on $\left(\Sigma_{n-1}, \bar{X}_{n-1}\right)$ and X_n, see Equation (2.6). So, the parameter space $\mathcal{Y}$ is $\left(\mathbb{R}^{d\times d}, \mathbb{R}^d\right)$.

Remark 3. *In step 5, we give one scheme to scale the variance of proposal distribution. The idea is that if the k-step average of acceptance rates is too large which implies that the jump scalar is too small, the proposal variance is required to be larger for the efficiency; if* $\alpha^{(k)}_{[n/k]}$ *is too small which implies that the jump scalar is too large, the proposal variance is required to be smaller for the efficiency. Here, we increase the proposal variance if* $\alpha^{(k)}_{[n/k]} > 0.3$*, and decrease it if* $\alpha^{(k)}_{[n/k]} < 0.3$*. Actually, the pair parameter* $(0.3, 0.3)$ *can be tuned. E.g. define* $\theta_n = \mathbf{1}(\alpha^{(k)}_{[n/k]} > 0.5)\exp(2d(\alpha^{(k)}_{[n/k]} - 0.5)) + \mathbf{1}(\alpha^{(k)}_{[n/k]} < 0.2)\exp(2d(\alpha^{(k)}_{[n/k]} - 0.2)) + \mathbf{1}(0.2 \leq \alpha^{(k)}_{[n/k]} \leq 0.5)$.

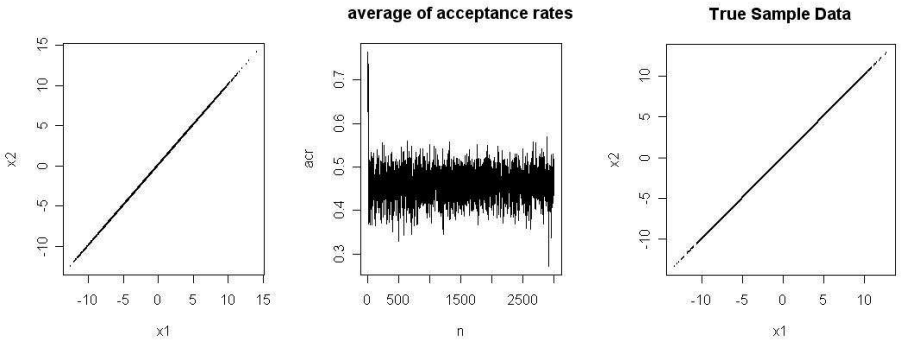

Figure 4.2: The left plot is the sample plot by running ADMG. The center plot is the 100-step average of acceptance rates. The right plot is true sample data.

Considering again the example in Section 3, we run ADMG by $300,000$ iterations, see the first two plots in Figure 4.2. The simulated data span roughly from $(-15, -15)$ to $(15, 15)$ which ADMG detects the target fasterly than MG and AM. The 100-step average $\{\alpha^{(100)}_n\}$ of acceptance rates is between 0.35 to 0.52. The last plot in Figure 4.2 is a true sample data from $t(\cdot)$. Comparing the first and last plot, ADMG exactly discovered the target region.

Remark 4. *From the discussion of the toy example, it is not difficult to find that when the target distribution is mainly supported on a long narrow region and it is highly correlated, ADMG is more efficient than the MG sampler and AM. In the high dimensional space, the phenomenon is more explicit.*

4.2 High dimensional Gaussian Needle

Here, we simulate a 10-dimensional Gaussian distribution on a long needle. Consider a 10-dimensional i.i.d. multivariate normal distribution $t'(x) \propto \exp\left(-x^\top D^{-1}x/2\right)$ where

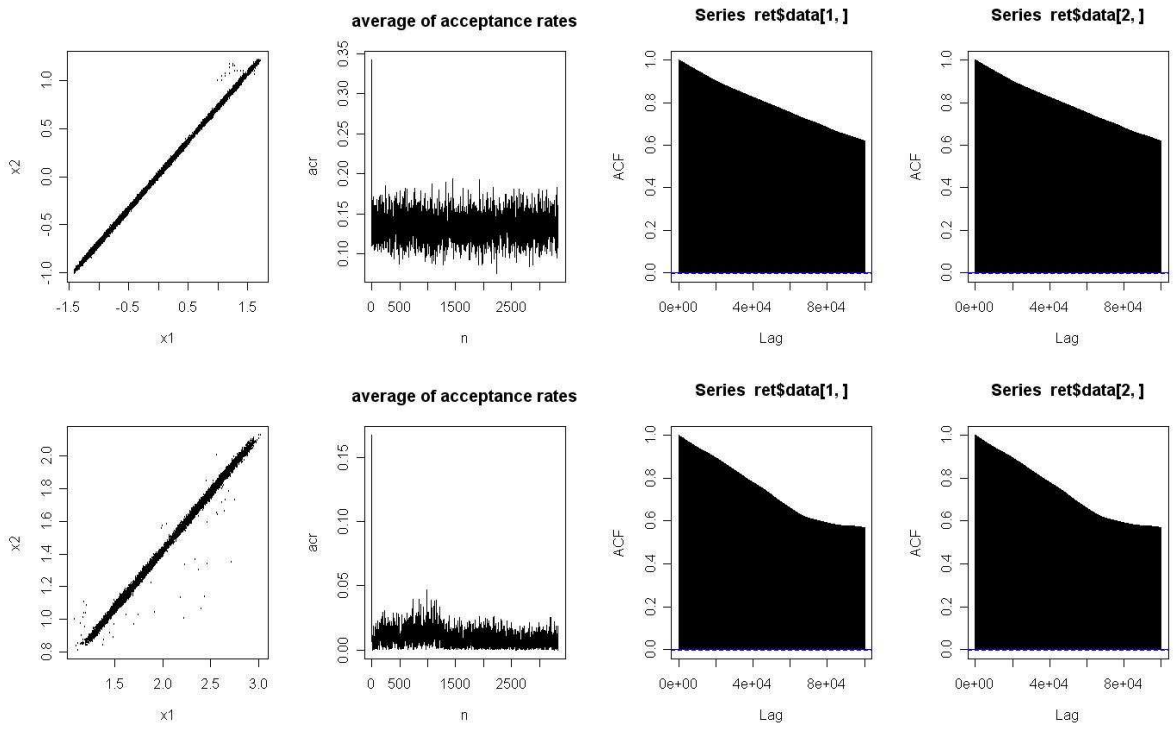

Figure 4.3: The top first plot is the sample plot by running MG. The top second plot is the 300-step average of acceptance rates. The top last two plots are the ACFs of the MG variables x_1 and x_2 with lag $100,000$. The bottom first plot is the sample plot by running AM. The bottom second plot is the 300-step average of acceptance rates. The bottom last two plots are the ACFs of the AM variables x_1 and x_2 with lag $100,000$.

$D = diag(20, 0.0001, \cdots, 0.0001)$ and $x \in \mathbb{R}^{10}$. We rotate by 45° the coordinate sequentially on the marginal plans $x_1 \perp x_2, \cdots, x_9 \perp x_{10}$. The corresponding transformations are $Q_{1,2}(45^\circ), \cdots, Q_{9,10}(45^\circ)$ where

$$Q_{i,j}(\theta) = I_{10} + \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cos\theta - 1 & 0 & \cdots & 0 & -\sin\theta & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sin\theta & 0 & \cdots & 0 & \cos\theta - 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

$$\begin{matrix} & & & i & & & & j & & & \end{matrix}$$

Thus, the interesting target density is

$$t(x) \propto \exp\left(-x^\top \left(QDQ^\top\right)^{-1} x/2\right), \tag{4.13}$$

where $Q = Q_{9,10}(45^\circ) \cdots Q_{1,2}(45^\circ)$.

We perform by $1,000,000$ iterations MG and AM algorithms where the initial point $X_0 \sim N(\overrightarrow{0}, diag(1, \cdots, 1))$. Figure 4.3 presents the sample data on the plane $x_1 \perp x_2$, and 300-step average acceptances of both algorithms. Both do stick in the quite short stripe. One is between $(-1.5, -1)$ to $(1.8, 1.3)$ with the length around 4.02, another is between $(1.0, 0.8)$ to $(3.0, 2.2)$ with the length around 5. Their lengths of the needle are far less than 35.78. Their estimates of autocorrelation functions (ACF) also show that the sample data have strong correlations.

We preform $1,000,000$ iterations using ADSSMG where the initial point has the same distribution as that of MG and AM. Figure 4.4 shows the sample data on the plane $x_1 \perp x_2$, the 300-step average of acceptance rates and the ACFs of ADSSMG variables x_1 and x_2 generated from ADSSMG. From these graphs, ADSSMG broadly detect the target with the narrow stripe roughly between $(-12, -10)$ to $(14, 10)$ with the length around 32.8. The average acceptance rate is roughly between 0.27 and 0.42. The ACFs of x_1 and x_2 almost tends to zero.

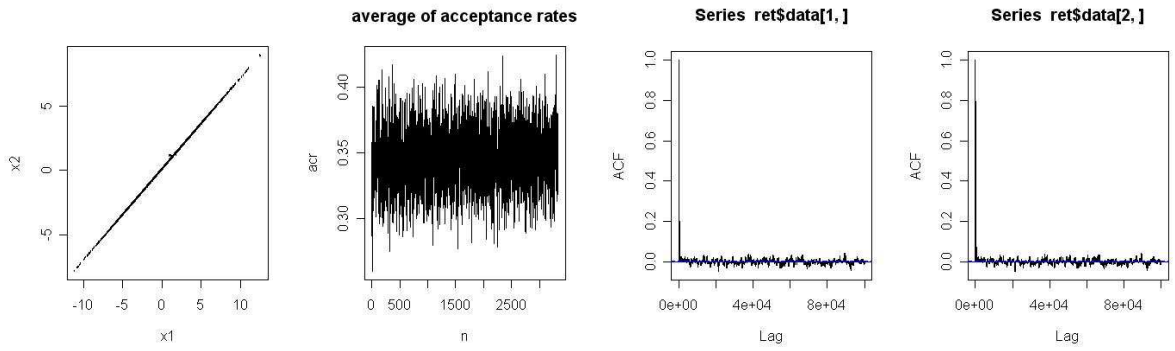

Figure 4.4: The first plot is the sample plot by running ADSSMG. The second plot is the 300-step average of acceptance rates. The right two plots are the ACFs of the ADSSMG variables x_1 and x_2 with lag $100,000$.

4.3 Ergodicity

In what follows, we shall write $n(z) := z/|z|$, and ∇ for the usual differential (gradient) operator.

Conditions for the target distribution and the proposal family:

A1: The target distribution $\pi(\cdot)$ on $\mathbb{R}^d$ is absolutely continuous w.r.t. Lebesgue measure μ_d with a continuously differentiable density t bounded away from zero and infinity on compact sets.

A2: For $e \in S^{d-1}$, $q_{e,\gamma}(x, x+ze) = q_{e,\gamma}(x, x-ze) := q_{e,\gamma}(z)$ for $\gamma \in \mathcal{Y}$ and $z \in \mathbb{R}$.

A3: For $\gamma \in \mathcal{Y}$ and $e \in S^{d-1}$, there exist $\delta_{e,\gamma} > 0$ and $\epsilon_{e,\gamma} > 0$ such that $q_{e,\gamma}(z) \geq \epsilon_{e,\gamma}$, for $|z| \leq \delta_{e,\gamma}$.

A4: There is a $\beta > 0$ and δ and Δ such that $\frac{1}{\beta} \leq \delta < \Delta < \infty$ for any (x^j, γ^j) with $\lim_j |x^j| = \infty$ and $\{\gamma^j\} \subset \mathcal{Y}$, we may extract a subsequence $(\tilde{x}^j, \tilde{\gamma}^j)$ with the property that for all $z \in [\delta, \Delta]$,

$$\begin{aligned} \lim_j \sup_{e \in \tilde{A}(\tilde{x}^j)} \frac{\pi(\tilde{x}^j)}{\pi(\tilde{x}^j + ze)} &\leq \exp(-\beta z) \\ \lim_j \sup_{e \in \tilde{R}(\tilde{x}^j)} \frac{\pi(\tilde{x}^j + ze)}{\pi(\tilde{x}^j)} &\leq \exp(-\beta z), \end{aligned} \tag{4.14}$$

where

$$\tilde{R}(x) := \{e \in S^{d-1} : \langle n(x), e \rangle \geq \frac{1}{\sqrt{d}}\} \text{ and } \tilde{A}(x) := \{e \in S^{d-1} : \langle n(x), e \rangle \leq -\frac{1}{\sqrt{d}}\}; \tag{4.15}$$

Moreover,

$$\inf_{\gamma\in\mathcal{Y}}\inf_{a\in S^{d-1}}\int_{\delta}^{\Delta} zq_{a,\gamma}(z)dz > \frac{d}{\beta(e-1)}. \tag{4.16}$$

A5: The target density t has continuous first derivatives and satisfies $\limsup_{|x|\mapsto\infty}\langle n(x), m(x)\rangle < 0$, where $m(x) := \nabla t(x)/|\nabla t(x)|$.

Remark 5. *A1 and A3 are used to ensure each Metropolis-within-Gibbs sampler is irreducible and aperiodic. A2 represents the symmetric property of the proposal family. A5 means that the target density has the proper contour surface as $|x|$ is sufficiently large. The condition A4 is a little abstract. The value $\frac{1}{\sqrt{d}}$ in Equation (4.15) ensures that given $x \in \mathbb{R}^d$, for any orthogonal normal coordinates $\{e_1, \cdots, e_d\}$, i.e. $|e_i| = 1$ and $e_i \perp e_j$ for $i \neq j$, there exists e_{i_0} for $i_0 \in \{1, \cdots, d\}$ such that $e_{i_0} \in \tilde{R}(x) \cup \tilde{A}(x)$. Equation (4.14) means the tails of target density on certain hypercone dependent of the dimension d decays in the exponential rate. Equation (4.16) implies that the first moment of the proposal family on some hypercone has an uniformly low bound, see the explanations in Bai et al. (2008).*

Remark 6. *When the decaying rate of target density is lighter-than-exponential, the β in the condition A4 can be arbitrarily large even infinity, A4 is implied by which the proposal family has an uniformly lower density function, see Theorem 6.3 in Bai et al. (2008).*

Remark 7. *Those conditions required in Bai et al. (2008) for adaptive Metropolis-within-Gibbs algorithms are weaker than A1-A5. Their conditions only require that the tails on each axis decay in the exponential rates, because the Metropolis random walks of their algorithm are restricted only on the coordinates. Here, the algorithm involves the rotation so that the Metropolis random walks may be performed on any appropriate direction on which the exponentially decaying rate is needed.*

Remark 8. *The condition A4 implies that* $\limsup_{|x|\to\infty}\langle n(x), \nabla \log t(x)\rangle < 0$ *(exponentially tailed).*

Theorem 1. *For the target distribution and the proposal family satisfying A1-A5, either ADRSMG or ADSSMG is ergodic.*

See the proof in Section A.

Remark 9. *Bai et al. (2008) showed that adaptive random-scan Metropolis-within-Gibbs algorithm under the original coordinates is ergodic, actually ergodicity of which is equivalent to that of ADRSMG.*

5 A Real-life Cohort Study with the competing risks

The Cox (1972) proportional hazards model is routinely used for failure time data. Cox (1975) studied the partial likelihood methods, also see textbook Kalbfleisch and Prentice (2002). Accordingly, Prentice (1986) proposed the Case-Cohort design to efficiently analyze Cohort data when most observations are censored, i.e. the interest events occur with low frequency. For epidemiologic studies, the cohort may be very large under the previous assumption. Self and Prentice (1988) proved the asymptotic normal properties of the estimate $\hat{\beta}$ under certain regularity conditions by using a pseudo-likelihood. Wacholder et al. (1989) proposed a bootstrap estimate of the variance of $\hat{\beta}$. Similar estimates for the variance were derived by Lin and Ying (1993) and Barlow (1994). Pintilie et al. (2009) used a modified partial likelihood to accommodate the modeling of the hazard of subdistribution for a Case-Cohort study. They used the Jackknife method to find the estimate's covariance matrix.

These frequentist methods mainly try to find the optimal coefficient estimates of covariates such that the pseudo-likelihood reaches the maximum. Here we utilize the Bayesian method through simulating the posterior distribution of the coefficients of covariates, and compare three algorithms: MG, AM and ADSSMG.

Here we describe the model used in Pintilie et al. (2009).

5.1 The Model Description

The hazard rate is defined as

$$\lambda(t,x) = \lim_{h\to 0^+} P(t \le T \le t+h \mid T \ge t, x)/h = \lambda_0(t) r(t,x),$$

where T is the random failure time, $\lambda_0(t)$ is an unspecified baseline hazard function, and $r(t,x)$ is the relative risk function. Here, we also assume that

$$\lambda(t,x) = \lambda_0(t)\exp(Z(t)^\top \beta) \tag{5.17}$$

where $Z(t) = (Z_1(t), \cdots, Z_p(t))$.

Suppose that the data consist of observations on a random vector Y with the density function $f(y \mid \theta, \beta)$ where β is the parameter of interest and θ is the nuisance parameter of high or infinity dimension. Suppose that Y can be transformed into $a_1, b_1, \cdots, a_n, b_n$ and $a^{(j)} = (a_1, \cdots, a_j)$ and $b^{(j)} = (b_1, \cdots, b_j)$. Assume that the joint density function can be written as

$$\prod_{j=1}^{n} f(b_j \mid b^{(j-1)}, a^{(j-1)}, \theta, \beta) \prod_{j=1}^{n} f(a_j \mid b^{(j)}, a^{(j-1)}, \beta).$$

The second term is called the *partial likelihood* of β. It should be noted that

$$L(\beta) = \prod_{j=1}^{n} f(a_j \mid b^{(j)}, a^{(j-1)}, \beta). \tag{5.18}$$

Suppose that there are a set of ordered pair times $(t_1^*, t_1), \cdots, (t_n^*, t_n)$ where t_j^*s are the entry times ($< t_j$) and t_js ($t_1 < \cdots < t_n$) are the event observing times. The corresponding censor variables are defined as

$$C_j = \begin{cases} 1 & \text{when the event of interest was observed} \\ 2 & \text{when the competing risk event was observed} \\ 0 & \text{when no event was observed} \end{cases}. \tag{5.19}$$

The set

$$R(t) = \{i : t_i^* \leq t \leq t_i; C_i = 0 \text{ or } 1\} \cup \{i : t_i^* \leq t; C_i = 2\}, \tag{5.20}$$

is the set of items at risk of failure at time t^-, just prior to time t.

Consider the instantaneous failure interval $[t_j, t_j + dt_j)$. The j^{th} term in the partial likelihood Equation (5.18) is

$$L_j(\beta) = \frac{\lambda(t_j, x_j)dt_j}{\sum_{l=1}^{n} Y_l(t_j)\lambda(t_j, x_l)dt_j}, \tag{5.21}$$

where $Y_l(t)$ indicates that $l \in R(t)$. By the assumption of Equation (5.17), the modified partial likelihood at the time of occurrence and the competing risks events with a specific weight for the Case-Cohort study is

$$L^*(\beta; x) = \prod_{j=1}^{n} \frac{\mathbf{1}(C_j = 1)\exp(\beta^\top x_j)}{\sum_{r \in R(t_j)} w_{rj}\exp(\beta^\top x_r)}, \tag{5.22}$$

where the weights $w_{rj} = \frac{\hat{G}(t_j)}{\hat{G}(t_j \wedge t_r)}$, and $\hat{G}(t_j)$ is the Kaplan-Meier estimator for the probability of censoring, see Kaplan and Meier (1958). The set $R(t)$ represents the case and time-matched controls at the Cohort follow-up time t. The covariates x_i can be time-dependent on t_i.

Here, we choose a prior $\mu(\cdot)$ (can be flat) for the coefficient β. The target distribution (the posterior distribution) that we want to simulate is

$$t(\beta) \propto \mu(\beta)L^*(\beta; x). \tag{5.23}$$

5.2 The analysis of Hypoxia Study

In the study, 109 patients with cervical cancer were treated at a cancer center between the year 1994 to 2000. Meanwhile two cancer marker were done in the time of diagnosis:

Table 5.1: Hypoxia study: 10 records are extracted from dataset

age	hgb	tumsize	IFP	HP_5	pelvicln	resp	pelrec	disrec	survtime	stat	dftime
78	119	7	8	32.1428571	N	CR	N	N	6.152	0	6.152
69	131	2	8.2	2.173913	N	CR	N	N	8.008	0	8.008
55	126	10	8.6	52.3255814	N	NR	Y	N	0.621	1	0.003
55	141	8	3.3	3.2608696	N	CR	Y	Y	1.12	1	1.073
50	95	8	18.5	85.4304636	Y	NR	Y	N	1.292	1	0.003
57	132	8	20	19.3548387	N	CR	N	N	7.929	0	7.929
53	127	4	21.8	44.5783133	E	CR	N	N	8.454	0	8.454
62	142	5	31.6	59.6774194	N	CR	Y	Y	7.116	0	7.107
23	145	5	16.5	29.1666667	N	CR	N	N	8.378	0	8.378
57	142	3	31.5	85.7142857	N	CR	N	N	8.178	0	8.178
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

hgb	Haemoglobin (g/l)
pelvicln	Pelvic node involvment: N=Negative, E=Equivoval,Y=Positive
pelrec	Pelvic disease observed: Y=Yes, N=No
disrec	Distant disease observed: Y=Yes, N=No
stat	Status at last follow-up: 0=Alive, 1=Dead

a hypoxia marker (HP_5) and the interstitial fluid pressure (IFP). HP_5 are defined as the percentage that a tumor had the oxygen level less than 5 millimetres of mercury (mmHg). IFP are measured at a number of locations in the tumor and a mean value per patients was calculated. There are totally six diagnosis variables (age, hgb, tumersize, IFP, HP_5, pelvicln) and five outcome variables (resp, pelrec, disrec, survtime, stat), see Table 5.1. The outcome variables include the information of the treatment, relapse and death. The response to treatment has two cases: complete response (CR) when the tumor has completely disappeared after treatment, and no response (NR) when either the disease has progressed to other sites or the tumor has not disappeared. Under the situation that resp is NR, if disease progressed distantly then disrec=Y; if the tumor still presents then pelrec=Y, see other analysis about this case in textbook Pintilie (2006).

Consider the modified partial likelihood Equation (5.22). Here the number n of observations is 109. We use all the diagnosis variables as the covariates so the β is defined on $\mathbb{R}^6$ where the components are sequentially age, hgb, tumsize, IFP, HP_5 and pelvicln. All the entry times t_i^*s are zero, and the failure times t_js are from the variable dftime. We use the outcome variables to define the censor variables C_j for competing risks,

$$C_j = \mathbf{1}(pelrec_j = Y) + 2\mathbf{1}(pelrec_j = N, stat_j = 1), \tag{5.24}$$

which means that the competing risk here is defined as that patients are dead and the tumors has disappeared.

Here we apply the MG, AM and ADSSMG to sample the data for the posterior distribution $t(\cdot)$ in Equation (5.23). We compare the estimates generated by three algorithms with the R package cmprsk - CRR. Table 5.2 shows the coefficients estimate generated by CRR, AM, MG, and ADSSMG. The three algorithms present very well. From Table 5.3, the standard errors of the coefficients generated by CRR and ADSSMG which show that the two

Table 5.2: The coefficient estimates by CRR, MG, AM and ADSSMG

	β_{age}	β_{hgb}	β_{tumsize}	β_{ifp}	β_{hp5}	$\beta_{\text{pelv.}}$
CRR	-0.025950	-0.013330	0.258900	0.031370	0.001198	0.497400
AM	-0.026309	-0.014401	0.245710	0.031485	0.001299	0.513099
MG	-0.026543	-0.013669	0.257617	0.031522	0.001398	0.506934
ADSSMG	-0.026521	-0.013658	0.256224	0.031679	0.001285	0.510447

Table 5.3: The standard errors by CRR and ADSSMG

	β_{age}	β_{hgb}	β_{tumsize}	β_{ifp}	β_{hp5}	$\beta_{\text{pelv.}}$
CRR	0.01564	0.01201	0.10690	0.01705	0.00633	0.33520
ADSSMG	0.01522	0.01298	0.10591	0.01982	0.00704	0.30021

groups of data are roughly same. Figure 5.6 presents the histograms of the sample marginal densities of HP_5 and IPF where the densities by ADSSMG are more normal than the other two. From Figure 5.5, AM's 100-step average of acceptance rates is relatively smaller than other two so that AM is relatively less efficient on the high dimensional space.

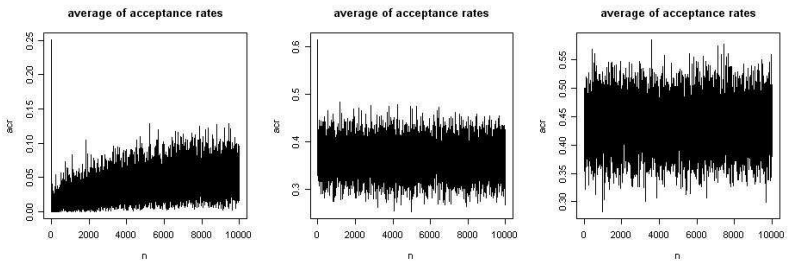

Figure 5.5: The left plot is the 100-step average of acceptance rates generated by AM; the center plot is the 100-step average of acceptance rates generated by MG; the right plot is the 100-step average of acceptance rates generated by ADSSMG.

Acknowledgement

The author is very grateful to Professor Jeffrey S. Rosenthal for his suggestions and guidance. The author also thanks Melania Pintilie for the useful discussions about the Cohort study and providing the real-life data.

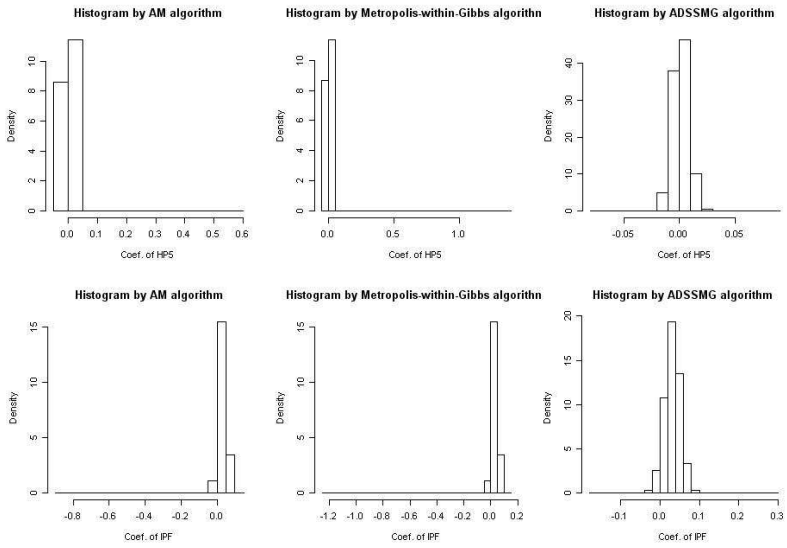

Figure 5.6: The top left is the histogram of HP_5 by AM; the top center is the histogram of HP_5 by MG; the middle right is the histogram of HP_5 by ADSSMG; the bottom left is the histogram of IPF by AM; the bottom center is the histogram of IPF by MG; the bottom right is the histogram of IPF by ADSSMG.

A Proof of Theorem 1

Lemma 1. *Suppose that the conditions A1-A4 are satisfied. Then there exists some* $s \in (0,1)$ *such that*

$$\limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} P_{DRS,\gamma} t^{-s}(x)/t^{-s}(x) < 1, \tag{A.25}$$

and

$$\limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} P_{DSS,\gamma} t^{-s}(x)/t^{-s}(x) < 1. \tag{A.26}$$

Proof: The proof for the case ADRSMG is omitted, because the procedure of Theorem 6.2 in Bai et al. (2008) can be adapted. We only discuss the ergodicity of ADSSMG.
Denote $V_s(x) = ct^{-s}(x)$ for some positive c such that $V_s(x) \geq 1$. Denote by $E^{(\gamma)}$ an array of d mutually orthogonal normal unit vectors for $\gamma \in \mathcal{Y}$. The array $E^{(\gamma)}$ is from the step 5 in the ADMG algorithm. From Equation (4.12),

$$P_{DSS,\gamma} V_s(x)/V_s(x) = \prod_{i=1}^{d} P_{E_i^{(\gamma)}} V_s(x),$$

and for each i,

$$P_{E_i^{(\gamma)}} V_s(x) \leq r(s) V_s(x),$$

where $r(s) := 1 + s(1-s)^{1/s-1}$.
Now, we assume that for any $s \in (0,1)$, $\limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} P_{\mathrm{DSS},\gamma} V_s(x)/V_s(x) \geq 1$. Then, there exists a sequence pair $\{(x^j, \gamma^j)\}$ with $\lim_{j\to\infty} |x^j| \to \infty$ and $\{\gamma^j\} \subset \mathcal{Y}$ such that $\lim_{j\to\infty} P_{\mathrm{DSS},\gamma_j} V_s(x^j)/V_s(x^j) \geq 1$.
Under the condition A4, we may extract from the sequence (x^j, γ^j) a subsequence $(\tilde{x}^j, \tilde{\gamma}^j)$ such that Equation (4.14) and (4.16) are satisfied. By the definition of $\tilde{R}(x)$ and the orthogonality of $E^{(\tilde{\gamma}^j)}$, there exists $i := i(\tilde{x}^j) \in \{1, \cdots, d\}$ with the element $E_i^{(\tilde{\gamma}^j)} \in E^{(\tilde{\gamma}^j)}$ such that $E_i^{(\tilde{\gamma}^j)} \in \tilde{R}(\tilde{x}^j)$ or $E_i^{(\tilde{\gamma}^j)} \in \tilde{A}(\tilde{x}^j)$ where $\tilde{A}(\tilde{x}^j)$ and $\tilde{R}(\tilde{x}^j)$ are defined in Equation (4.15). Without loss of generalization, we write $E_i^{(\tilde{\gamma}^j)} \in \tilde{R}(\tilde{x}^j)$.
So,

$$P_{\mathrm{DSS},\gamma^j} V_s(x^j)/V_s(x^j) \leq r^{d-1}(s) P_{E_i^{(\gamma^j)}} V_s(x^j)/V_s(x^j). \tag{A.27}$$

Adapting the techniques of the proof in Theorem 6.2 in Bai et al. (2008), by Equation (4.14), we have

$$\begin{aligned} & P_{E_i^{(\tilde{\gamma}^j)}} V_s(\tilde{x}^j)/V_s(\tilde{x}^j) \\ \leq\ & r(s)(1 - 2K_{i,\tilde{\gamma}^j}(0)) + K_{i,\tilde{\gamma}^j}(\beta s) + K_{i,\tilde{\gamma}^j}(0) + K_{i,\tilde{\gamma}^j}(\beta(1-s)) - K_{i,\tilde{\gamma}^j}(\beta), \end{aligned}$$

where $K_{i,\gamma}(t) = \int_{\delta\vee 1/\beta}^{\Delta} e^{-tz} q_{E_i^{(\gamma)},\gamma}(z)dz$.
Define $H_{i,\gamma}(\beta, s) := r^{d-1}(s)\left[r(s)(1 - 2K_{i,\gamma}(0)) + K_{i,\gamma}(\beta s) + K_{i,\gamma}(0) + K_{i,\gamma}(\beta(1-s)) - K_{i,\gamma}(\beta)\right]$.
Hence,

$$P_{\mathrm{DSS},\tilde{\gamma}^j} V_s(\tilde{x}^j)/V_s(\tilde{x}^j) \leq H_{i,\tilde{\gamma}^j}(\beta, s). \tag{A.28}$$

Thus,

$$\begin{aligned} H_{i,\tilde{\gamma}^j}(\beta, 0) &= 1; \\ \frac{\partial H_{i,\tilde{\gamma}^j}}{\partial s}(\beta, 0) &= (d - 2K_{i,\tilde{\gamma}^j}(0))r'(0) - \beta \int_{\delta}^{\Delta} z q_{E_i^{(\tilde{\gamma}^j)},\tilde{\gamma}^j}(z)\mu(dz) + \beta \int_{\delta}^{\Delta} z e^{-\beta z} q_{E_i^{(\tilde{\gamma}^j)},\tilde{\gamma}^j}(z)\mu(dz) \\ &\leq d/e - \beta(1 - 1/e) \int_{\delta}^{\Delta} q_{E_i^{(\tilde{\gamma}^j)},\tilde{\gamma}^j}(z)dz. \end{aligned}$$

By Equation (4.16),

$$\limsup_{j\to\infty} \frac{\partial H_{i,\tilde{\gamma}^j}}{\partial s}(\beta, 0) < 0.$$

Therefore, $\limsup_j H_{i,\tilde{\gamma}^j}(\beta, s) < 1$ for some $s \in (0,1)$, which leads to a contradiction. □

Proof of Theorem 1: By Lemma 1 and Proposition 2.3 in Bai et al. (2008), Containment holds. From Proposition 5.9 in Bai et al. (2008), by the conditions A5, Diminishing adaptation holds. Therefore, by Theorem 13 in Roberts and Rosenthal (2007), the two algorithms are ergodic. □

References

H.C. Andersen and P. Diaconis. Hit and run as a unifying device. *Journal de la Société Française de Statistique*, 148(5):5–28, 2007.

C Andrieu and E Moulines. On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006.

C. Andrieu and C.P. Robert. Controlled mcmc for optimal sampling. *Preprint*, 2002.

Y.F. Atchadé and G. Fort. Limit Theorems for some adaptive MCMC algorithms with subgeometric kernels. *Preprint*, 2008.

Y.F. Atchadé and J.S. Rosenthal. On Adaptive Markov Chain Monte Carlo Algorithms. *Bernoulli*, 11(5):815–828, 2005.

Y. Bai. Simultaneous drift conditions on adaptive Markov Chain Monte Carlo algorithms. *Technical Report in Department of Statistics at the University of Toronto*, 2009.

Y. Bai, G.O. Roberts, and J.S. Rosenthal. On the containment condition of adaptive Markov Chain Monte Carlo algorithms. *Technical Report in Department of Statistics at the University of Toronto*, 2008.

W.E. Barlow. Robust variance estimation for the case-cohort desing. *Biometrics*, 50:1064–1072, 1994.

M. Bèdard and D.A.S. Fraser. On a directionally adjusted metropolis-hastings algorithm. *Preprint*, 2008.

C.J.P. Bélisle, H.E. Romeijn, and R.L. Smith. Hit-and-run algorithms for generating multivariate distributions. *Math. of Operation. Research*, 18(2), 1993.

A.E. Brockwell and J.B. Kadane. Indentification of regeneration times in mcmc simulation, with application to adaptive schemes. *J. Comp. Graph. Stat*, 14:436–458, 2005.

M.H. Chen and B.W. Schmeiser. Performance of the gibbs, hit-and-run, and metropolis samplers. *J. Comp. and Graph. Stats.*, 2(3):251–272, 1993.

M.H. Chen and B.W. Schmeiser. General hit-and-run monte carlo smapling for evaluating multideimensional integrals. *Operation Research letter*, 19:161–169, 1996.

D.R. Cox. Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972.

D.R. Cox. Partial likelihood. *Biometrika*, 62:269–272, 1975.

R.V. Craiu, J.S. Rosenthal, and C. Yang. Learning from thy neighbor: Parallel-chain adaptive mcmc. *Preprint*, 2008.

G. Fort, E. Moulines, G.O. Roberts, and J.S. Rosenthal. On the geometric ergodicity of hybrid samplers. *J. Appl. Prob.*, 40:123–146, 2003.

W.R. Gilks, G.O. Roberts, and E.I. George. Adaptive direction sampling. *The statistician*, 43:179–189, 1994.

W.R. Gilks, G.O. Roberts, and S.K. Sahu. Adaptive markov chain monte carlo. *J. Amer. Statist. Assoc.*, 93:1045–1054, 1998.

H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7: 223–242, 2001.

W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series, 2002.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481, 1958.

D.E. Kaufman and R.L. Smith. Direction choice for accelerated convergence in hit-and-run sampling. *Operations Research*, 46(1), 1998.

D.Y. Lin and Z. Ying. Cox regression with incomplete covariate measurements. *J. Amer. Statist. Assoc.*, 88:1341–1349, 1993.

L. Lovász. Hit-and-run mixes fast. *Math. Program.*, 86(Ser. A):443–461, 1999.

L. Lovász and S. Vempala. Hit-and-run is fast and fun. *preprint, Microsoft Research*, 2003.

L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM J. Comput.*, 35:985–1005, 2006.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phsy.*, 21:1087–1091, 1953.

M. Pintilie. *Competing Risks: A Practice Perspective*. Wiley Series, 2006.

M. Pintilie, Y. Bai, L.S. Yun and D. Hodgson. The analysis of case cohort design in the presence of competing risks with application to the analysis of the effect of treatment for hodgkin lymphoma on cardiac events. *preprint*, 2009.

R.L. Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73:1–11, 1986.

G.O. Roberts and W.R. Gilks. Convergence of adaptive direction sampling. *J. Multiv. Analys.*, 49:287–298, 1994.

G.O. Roberts and J.S. Rosenthal. Examples of Adaptive MCMC. *J. Comp. Graph. Stat., to appear*, 2006a.

G.O. Roberts and J.S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and Trans-dimensional Markov Chain. *Ann. Appl. Prob.*, 16(4):2123–2139, 2006b.

G.O. Roberts and J.S. Rosenthal. Coupling and Ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Prob.*, 44:458–475, 2007.

E. Saksman and M. Vihola. On the Ergodicity of the Adaptive Metropolis Algorithms on Unbounded Domains. *Preprint*, 2008.

S.G. Self and R.L. Prentice. Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.*, 16:64–81, 1988.

S. Wacholder, M.H. Gail, D. Pee, and R. Brookmeyer. Alternative variance and efficiency calculations for the case-cohort design. *Biometrika*, 76:117–123, 1989.

C. Yang. Recurrent and Ergodic Properties of Adaptive MCMC. *Preprint*, 2008.